IN THIS ISSUE
4 Papers from IASSIST 2010

ON PAGE 5
Editor Karsten Boye Rasmussen

ON PAGE 26
Membership Information
In this issue

5 Editor's Notes
Karsten Boye Rasmussen

6 Social Science Data Archiving and Needs of the Public Sector:
the case of Estonia by Rein Murakas and Andu Rämmer

9 Wikiprogress and Wikigender
A way forward for online collaboration by Angela Hariche, Estelle Loiseau and Philippa Lysaght

15 Secure Data Service:
An improved access to disclosive data by Reza Afkhami, Melanie Wright, Mus Ahmet

19 A User-Driven and Flexible Procedure for Data Linking
by Cees van der Eijk and Eliyahu V. Sapir
Editor’s notes

New format - new papers
Walter Piovesan – our publication officer – had a biking accident. To show that nothing is so bad that it is not good for something, Walter used his recovery time to redesign the IQ. I hope you like it. I like it! Furthermore, Walter is part of the local arrangements committee, along with Mary Luebbe, in charge of the coming 2011 IASSIST conference, so he is a busy guy. And I’m happy to say that Walter is now fully recovered.

Welcome to the IASSIST Quarterly volume 34 (2) of 2010.

This issue of the IQ features the following papers:

Rein Murakas and Andu Rämmer from the Estonian Social Science Data Archive (ESSDA) at the University of Tartu describe in their paper “Social Science Data Archiving and Needs of the Public Sector: the Case of Estonia” how the archive had a historical background in the empirical research of the Soviet Union. After Estonian independence the creation of the social science data bank acquired momentum and the data archive was established in 1996. The following year the data archive became a member of the European Council of Social Science Data Archives (CESSDA). This paper was presented at the IASSIST 2009 conference in the session “Building Data Archives and User Communities: Greece, Estonia and Ethiopia”.

From the historical background we move to web 2.0 in a paper by Angela Hariche, Estelle Loiseau and Philippa Lysaght on “Wikiprogress and Wikigender: a way forward for online collaboration”. The paper was presented at the IASSIST 2010 Conference at Cornell University, Ithaca. The authors are working at the OECD and the paper’s statement is that “collaborative platforms such as wikis along with advances in data visualisation are a way forward for the collection, analysis and dissemination of data across countries and societies”, and their paper “presents two wiki platforms that can support such a vision: Wikiprogress and Wikigender”. These applications support the OECD’s shift from measuring (only) economic production to measuring well-being and the general progress of societies.

The third paper addresses an issue of central importance for most data archives. The question concerns balancing data confidentiality and the legitimate requirements of data users. This is a key problem of the Secure Data Service (SDS) at the UK Data Archive, University of Essex. The paper “Secure Data Service: an improved access to disclosive data” by Reza Afkhami, Melanie Wright, and Mus Ahmet was presented at IASSIST 2010 in the session “Secure remote access to restricted data”. The SDS will allow researchers remote access to secure servers at the UK Data Archive. In the paper many related issues in legal, contractual, educational, and technical areas are discussed and examined. On top of this support the SDS may potentially open access to previously closed data and thus further facilitate new research.

The last article has the title “A user-driven and flexible procedure for data linking”. The authors are Cees van der Eijk and Eliyahu V. Sapir from the Methods and Data Institute at the University of Nottingham. The paper was presented at IASSIST 2010 in the panel on “Virtual Research Environments: Tools for Presenting and Storing Data”. The data linking relates to research combining several different datasets. The paper defines the relational database operations that can be used as flexible and end-user directed choices. The implementation is developed for the PIREDEU project in comparative electoral research. The authors are combining traditional survey data with data from party manifestos and state-level data. The paper exemplifies the linking of voter and manifesto studies, and also addresses more complex data linking and merging.

Articles for the IQ are always very welcome. They can be papers from IASSIST or other conferences, from local presentations or papers especially written for the IQ. If you don’t have anything to offer right now, then please prepare yourself for the next IASSIST conference and start planning for participation in a session there. Chairing a conference session with the purpose of aggregating and integrating papers for a special issue IQ is much appreciated as the information reaches many more people than the session participants and will be readily available on the IASSIST website at http://www.iassistdata.org.

Authors are very welcome to take a look at the description for layout and sending papers to the IQ: http://iassistdata.org/iq/instructions-authors.

Authors can also contact me via e-mail: kbr@sam.sdu.dk. Should you be interested in compiling a special issue for the IQ as guest editor or editors I will also be delighted to hear from you.

Karsten Boye Rasmussen
March 2011
Editor
Social Science Data Archiving and Needs of the Public Sector: the Case of Estonia by Rein Murakas and Andu Rämmer

Abstract
Estonia was one of the pioneering empirical data research centres in the Soviet Union since the 1960s, with computing centres located at the largest research institutions. After the restoration of Estonian independence, a team of social scientists developed an initiative to save unique data produced in the preceding decades. Higher Education Support Program from Open Society Foundations supported early data conversion projects. Establishment of the Estonian Social Science Data Archive in 1996 at the University of Tartu provided a focal point for data acquisition and support efforts, but its activities have been constrained by lack of funding and the size of the Estonian social science research community. ESSDA has broadened its user base by offering services to employees in the public sector and teaching general principles of data use and archiving to social science students.

Development of social science and data archiving in Estonia before founding of ESSDA
Estonia was incorporated into the Soviet Union only in 1940 after a period of independence. For that reason, it had a relatively open ideological atmosphere and many of its citizens spoke foreign languages. For these reasons, new ideas were accepted more easily than in the rest of the Soviet Union. Thus it was not surprising that Estonia became (along with the Russian cities of Leningrad and Moscow) one of the pioneering empirical data research centres from the 1960s (Titma 2002). Two of the largest mainframe computing centres were housed by Estonian Radio and the University of Tartu.

The research environment in Soviet period produced relatively good quality data products in robust computing environments. However, shortcomings in university education in sociological disciplines caused data analysis methodologies to be weak. Improvements in sociology education in universities started only in 1989 when the Soviet Union was very near to its collapse. That milestone sharply accelerated the development of quantitative methodologies and expanded the number of potential secondary data users in Estonia.

After the restoration of Estonian independence, a team of social scientists from the University of Tartu made up an initiative group in 1993 that had two primary goals: to create a social sciences data bank and to develop a strategy for saving and encouraging use of the research materials collected by the Estonian social scientists during the previous decades. Their application for support to the Higher Education Support Program (HESP) was successful; a grant gave a chance to “save” unique data from previous decades.

In 1996, the Estonia Social Science Data Archive was officially established as an interdisciplinary centre of the Faculty of Social Sciences of Tartu University and as a national social science data bank. At the end of that year, ESSDA hosted an International conference on data archives and their functions in social research in Eastern Europe (Murakas and Rämmer 2001). ESSDA became a full member of the European Council of Social Science Data Archives (CESSDA) in 1997.

Developments of data archiving in Estonia after the establishment of ESSDA
Lack of the funding seriously restricted development of the Estonian Social Science Data Archive in the late 1990s (Murakas and Rämmer 2002). The situation was quite critical and the main goal for ESSDA in that period was survival. Several applications for funding were rejected, so the Faculty of Social Sciences of the University of Tartu has been our only regular funding source. The archive was also supported by the national programme Collections for the
Humanities and Natural Sciences” in 2005–2008, especially in activities connected with Soviet-time data. Despite very limited resources, ESSDA has continued its activities compiling new data and organizing existing data. Our electronic data collections were supplemented with complementary non-electronic archives of Soviet period studies (study reports, unpublished analyses) from Estonian Radio and the former Laboratory of Educational Studies (that continued its activities as Department of Sociology) at University of Tartu.

The number of users of the ESSDA has increased. In addition to Estonian users, some are foreign social scientists and journalists. A growing number of international publications (Kaplan 2006, Kaplan and Brady 2009, Brady and Kaplan 2011) are based on the archived Estonian data, ESSDA itself is noted as an important institution in various overviews about the state of Estonian social science (Kroos, Murakas and Veski 2009, Murakas and Rämm 2000, Rosimannus and Titma 2004, Titma 2002). At the same time interest among Estonian researchers in studies housed by foreign archives is growing. Most are interested in Eurobarometer surveys. Martiotti and Stefanizzi (2000) stressed the importance of Eurobarometer as one of the most comprehensive and continuous academic survey programs made available to the academic community thanks to the data archiving.

ESSDA has developed some important initiatives. Accompanying printed questionnaires of Soviet-era datasets were scanned and now make up an essential part of our online database. We tried to launch a regular international electronic journal “Estonian Social Science Online”. Due to lack of finances, only two first issues in English were launched. However, two subsequent numbers were issued with articles in Estonian. These featured the presentations of the Estonian Annual Conferences of Social Sciences.

ESSDA also faces many challenges. The biggest limitation that emerged in acquiring new data is the lack of trust. Like most spheres of life in post-communist societies (Howard 2003) the lack of trust also characterizes potential providers of social data. For example, they can be worried that local customers at Tartu University may enjoy advantages of access to the data not available to others. It appears that the tradition of sharing data among social scientists is remarkably weak and depends on personal relationships. Secondly, support from the public sector is often cooperative but comes with no financial assistance. Thirdly, the permanent lack of funding generally reduces chances of finding additional means of financing. Often the work of the data archive was regarded as some small local initiative of enthusiasts and it was not financed from state level structures.

In the framework of structural reform at Tartu University ESSDA has been re-organized to a consortium in 2008. ESSDA’s statute was also renewed. This organizational reform would probably give us better possibilities for direct cooperation with different institutions involved with using and producing of social science data.

Current data sources and data usage in Estonia

Estonia is one of the smallest EU countries with a total population of 1.34 million at January 1st, 2010. Although membership in the EU boosted international cooperation among social scientists (Murakas et al. 2007), the total number of social scientists in Estonia is one of smallest among member countries. On the basis of very discussable statistical data, in Estonia in 2009 947 people are working as social science researchers in non-profit institutions. In full-time equivalents the same number is 474. Those data are based on broad definition of social science. If we exclude law and economy professionals, not directly connected with social science survey data stored in ESSDA, we can say, that only a few hundred social scientist are potential users of our archive data. They are organised into 30-40 small research groups that use and need data in different ways. If ESSDA data collections consist of only local data, they are not a good resource base for publications in international peer-reviewed journals, the main criterion for evaluating academic contributions among the social sciences in Estonia.

The Ministry of Education and Research together with the Estonian Academy of Sciences launched a process of compiling the Estonian Research Infrastructures Roadmap. To develop possibilities of international data usage in social sciences, ESSDA made a proposal to include the Council of European Social Science Data Archives (CESSDA) as an important international infrastructure in the roadmap. ESSDA’s proposal was not accepted with suggestion to apply again in the next round. But it is important that from social sciences, the proposal “Estonia in European Social Survey” was accepted for the roadmap.

Archived Estonian data can be essential for the narrow branch of historical analysis of the former Soviet Union. For example we can mention Estonian Science Foundation grant “Changing cultural dispositions of Estonians through the four decades: from the 1970s to the present time” headed by professor Marju Lauristin, started in 2009. The historical part of the project is based on the re-analysis of the Soviet and transition-time Estonian data from ESSDA.

However, research groups use mostly public data from comparative surveys (like European Social Survey) that are available online, but also official statistics collected by Estonian Statistical Office that is stored in their online database and data collected by joint projects with international partners. However, we can conclude that there cannot be very high demand for data from ESSDA’s traditional clients for the foreseeable future. To enlarge the circle of potential customers, reorganization of the activities and change of paradigm were needed.

Raising the importance of alternative customers from the public sector

In 2009, the Estonian public sector employed about 159, 000 people, about 47% of them with higher education, Master’s or PhD degree. About 37, 000 people are working in the state-level institutions (branch Public administration and defence, compulsory social security in Estonian classification of economic activities). It is difficult to say how many of public sector employees can be considered analysts who need social science data to support their work but in any case their number is much higher than that of academic social scientists.

Fast economic development triggered new divisions in society, and the need to encourage use of social science research information by public sector staff is urgent. This information is of growing significance to Estonian public policy and its development programs in recent years. Based on surveys (Kasemets 2009), the real situation suggests that the use of social science data is rather modest in the public sphere. Possible rise in interest toward data resources should be connected with the need to raise the effectiveness of the public sector (especially due to budget cuts) and also with the realisation of different projects that are directed to the development of public sector. The number of projects financed from European structural funds that require increasing of quantitative methodologies based on social analyses has increased sharply after joining the European Union.

At present, the focus of sharing the fruits of applied research is production and distribution of research. These fulfill the legislative mandate that the results of applied research projects that are publicly funded by the Estonian government be widely distributed by government agencies. However, such reports are generally published only on web sites of these agencies. There is no coordination among agencies to provide a comprehensive view of public research investments, for example to establish centralized database of research reports.
To help address this, ESSDA submitted a proposal to establish an inter-agency metainformation database with information on these dispersed research projects. However, the proposal was not funded because of emerging institutional barriers. Every institution needs for its purposes specific information, for example, an overview about research projects connected with higher education or multicultural issues. As a result, the current solution is that ESSDA started establishing metainformation database by topics depending on the available funding sources. This means that single metainformation database is accessible via different user interfaces via different user interfaces like separate interface for educational or for cultural surveys. In 2009, ESSDA already developed all-Estonian survey information database about higher education and also small database about multicultural problems like for educational surveys or for cultural surveys. Plans for the future are to build a large NESSTAR-based meta-information data base that will cover all research areas. Our NESSTAR server is launched and the number of resources in our virtual library is increasing. Early data descriptions are in Estonian as we try to extend the NESSTAR community in the Estonian public sector.

However, even if research reports are sometimes insufficient, there is currently little further interest from the public sector representatives to analyse data themselves. As a general rule, they do not need complex secondary analysis but quick answers to very concrete questions. In addition, they sometimes lack relevant skills and appropriate software expertise. These are difficult to address when public sector budgets are being cut. ESSDA may have a future role in providing a service to analyse and interpret survey results for many different research datasets to support public policy work. As part of the Faculty of Social Sciences, ESSDA could use graduate students as part-time analysts. In the short term, there is no funding for this. But in the longer term, the data service tradition can be institutionalised if it can become self-financed.

**Raising the profile with teaching and training**

A large percentage of social science students start their careers in the public sector after graduating from the university. As a result, working with students can supply them with the data use and archiving skills they may need when they enter the workforce. At the University of Tartu, almost all social science students (excluding law students) get at least basic information about social science data archives and learn fundamentals of secondary analysis in the introductory social science courses (“Introduction to Social Sciences”, Basics of Social Analysis”). We also teach advanced data archiving study courses (“Social Science Data Archives”, Databases in Social Sciences”) or sociology students on a regular basis and offer a special course “Re-contextualization of Sociological Inquiries from the Soviet Era” about data resources and interpretation difficulties of Soviet-period data to PhD students.

When our alumni attain key positions in the public sector, they are familiar with the potential of secondary analysis based on data archiving. They will also be more competent in using resources offered by the data archive.

Dangers associated with misuse of data and disregard of rules for data protection can be serious obstacles to accepting the products of data analysis (Jagodzinski and Moschner 2007). We initiated activities directed toward raising the knowledge and competency of the public sector in the field of social data analysis. To that end we developed specific components in special continuing vocational training (CVT) courses about the methodology and research design. We already have positive experiences about providing such training with education and criminal justice analysts, secondary school leaders etc. Conveying knowledge of the advantages and strengths of data libraries is not an easy job, but such problems are no unusual and characterize the current state of social sciences in general.

**References**


**Notes**

1. Rein Murakas and Andu Rämmer, University of Tartu, Estonia and Estonian Social Science Data Archive (ESSDA). Contact: Andu Rämmer, andu@ut.ee.
Abstract
In recent years there has been an explosion of web 2.0 technologies such as Facebook, Twitter, blogs and wikis. The world is becoming ever more connected with great technological strides which are increasingly allowing for more access to those who previously were disconnected. Web technologies are fostering mass collaboration on a scale that has never been seen before, allowing for a collective intelligence to develop on specific topics and themes. The crucial element to this participation is the openness of new web platforms, engaging everyone from policy makers to local NGOs to interested individuals. Recent technological advances are very timely as the world is also becoming more complex, and society has to face a number of global challenges that it can no longer address without connecting to others: from climate change to the financial crisis to sustainable livelihoods, these problems are global and therefore require global thinking. An important part of the problem-solving process involves the collection of information and data from diverse sources. This paper is concerned with the idea that collaborative platforms such as wikis along with advances in data visualisation are a way forward for the collection, analysis and dissemination of data across countries and societies, and presents two wiki platforms that can support such a vision: Wikiprogress and Wikigender. The paper first explains the wider movement behind the two wikis, namely the OECD-hosted Global Project on Measuring the Progress of Societies, then explores the importance of fostering collaboration and data dissemination through Wikiprogress and Wikigender. Finally, this paper concludes that collaboration is required for the shift from measuring economic production to measuring well-being to occur. Such collaboration is facilitated by web 2.0 technology.

Introduction
Societies are becoming increasingly connected thanks to advances in new technologies. Global problem-solving has been made easier through the new forms of collaboration this connectivity allows. A plethora of web 2.0 initiatives have flourished around the world, including issues like tackling climate change to solutions for sustainable development, the reduction of poverty and hunger, and decreasing the wage gap between women and men. The range of topics tackled on the Internet reflects societies’ increasing concern with their quality of life along with their material well-being. Leading economists studying the measurement of those factors which contribute to a person’s welfare conclude that “GDP is an inadequate metric to gauge well-being over time particularly in its economic, environmental, and social dimensions, some aspects of which are often referred to as sustainability.” Numerous initiatives worldwide, such as the OECD-hosted Global Project on Measuring the Progress of Societies, amongst others, have shown an interest in going beyond the mere economic prosperity of a nation – claiming that we should look beyond traditional measures such as Gross Domestic Product (GDP) in determining the well-being of societies.

Wikiprogress, the official platform of the Global Project on Measuring the Progress of Societies is a collaborative tool for measuring progress beyond GDP that exemplifies this new form of collaboration and the potential Web 2.0 technologies offer in solving global problems. It is important initially to understand the underlying movement and what “measuring progress” actually means. Examining the background of ‘measuring progress’ and identifying the key players will clarify the OECD’s choice of a ‘wiki’ platform and what impact this has on the measuring progress movement.

Box 1: Wikiprogress: goal and purpose

Goal: To create a web community around the vision of measuring the progress of societies by creating a place where progress data and research articles can be loaded, visualised and analysed so good decisions about societies can be made at the local, national and international levels.

Purpose: To invite and inform all parts of the progress community, citizens and policy makers alike to the debate on progress-related initiatives by creating a robust wiki of related research and statistics.
1. “Measuring Progress”: an overview of the movement

For the last century the most commonly used indicator of economic performance has been GDP, also used as an indicator of national performance. It is commonly assumed that GDP growth is synonymous with progress, that a rise in GDP mirrors an increase in the quality of life and well-being of a particular society. This, however, is not the case: GDP only measures what is ‘produced’ – it is a very narrow economic measure that fails to take into account other elements which are important in the daily lives of citizens such as security, pollution and health care. GDP calculations include economic activities that reduce well-being such as traffic congestion, which increases fuel consumption, or those which remedy the costs of economic growth like pollution abatement.

Currently there are many organisations who are looking at new measures of progress and calling for indicators going beyond GDP to measure the well-being of individuals, of our societies and our environment. A key message from the Stiglitz-Sen-Fitoussi Commission on Economic Performance and Social Progress in 2009 shows an urgency for developing such measures, declaring that now is the time for our measurement systems to shift the focus from measuring economic production to measuring people’s well-being. Furthermore, such measures of well-being should be considered and analysed within the context of sustainability.

The movement is not concerned solely with developing progress indicators but also the development of a collaborative community, working to determine how we measure progress. This kind of knowledge sharing and creation can aid statistical organisations or government by a nurturing community that encompasses diverse actors, including individuals worldwide. For this reason, the Global Project on Measuring the Progress of Societies chose to use a wiki platform, called Wikiprogress. The box below outlines the goal and

![My SIGI: Build Your Own Ranking](image-url)

Filter the ranking by region:
East Asia and Pacific | Europe and Central Asia | Latin America and the Caribbean | Middle East and North Africa | South Asia | Sub-Saharan Africa | All

Place checkmarks next to the indicators you wish to include in your gender index. Use the numbered buttons to adjust the weighting, 1 = not important, 5 = very important.

- Family Code
- Civil Liberties
- Physical Integrity
- Son Preference
- Ownership Rights

Only include countries with full data.

Calculate Ranking  Reset to Defaults

Map View

Your index at a glance: The map shows each country’s ranking in your index. Move your mouse over a country to see its rank in your index and in SIGI.

Figure 1. Visualisation of the www.my.genderindex.org tool
purpose of Wikiprogress, reflecting the importance of fostering a collaborative online community.

2. Why a wiki?
According to the world's largest online encyclopaedia, Wikipedia, a wiki is “a website that allows the easy creation and editing of any number of interlinked web pages via a web browser using a simplified mark-up language or a text editor.” In other words, a wiki is a type of web platform that allows user-generated content. The content in a wiki is created by a community of users working together to develop information on a particular subject or within a particular field. It is essentially a database of information that can be browsed, searched, created and edited.

<table>
<thead>
<tr>
<th>Country</th>
<th>Indicator</th>
<th>Time</th>
<th>2009</th>
<th></th>
<th></th>
<th>GPI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Afghanistan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albania</td>
<td></td>
<td></td>
<td></td>
<td>58.28</td>
<td>59.46</td>
<td>57.02</td>
</tr>
<tr>
<td>Algeria</td>
<td></td>
<td></td>
<td></td>
<td>23.4</td>
<td>23.36</td>
<td>23.44</td>
</tr>
<tr>
<td>Andorra</td>
<td></td>
<td></td>
<td></td>
<td>98.53</td>
<td>99.61</td>
<td>97.35</td>
</tr>
<tr>
<td>Angola</td>
<td></td>
<td></td>
<td></td>
<td>40.14</td>
<td>44.91</td>
<td>35.41</td>
</tr>
<tr>
<td>Anguilla</td>
<td></td>
<td></td>
<td></td>
<td>94.54</td>
<td>98.66</td>
<td>90.87</td>
</tr>
<tr>
<td>Antigua and Barbuda</td>
<td></td>
<td></td>
<td></td>
<td>64.8</td>
<td>62.29</td>
<td>67.41</td>
</tr>
<tr>
<td>Argentina</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Armenia</td>
<td></td>
<td></td>
<td></td>
<td>33.3</td>
<td>31</td>
<td>35.99</td>
</tr>
<tr>
<td>Aruba</td>
<td></td>
<td></td>
<td></td>
<td>104.6</td>
<td>106.81</td>
<td>102.34</td>
</tr>
<tr>
<td>Australia</td>
<td></td>
<td></td>
<td></td>
<td>82.29</td>
<td>83.22</td>
<td>81.3</td>
</tr>
<tr>
<td>Austria</td>
<td></td>
<td></td>
<td></td>
<td>95.11</td>
<td>95.34</td>
<td>94.86</td>
</tr>
<tr>
<td>Azerbaijan</td>
<td></td>
<td></td>
<td></td>
<td>26.48</td>
<td>26.05</td>
<td>26.99</td>
</tr>
<tr>
<td>Bahamas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bahrain</td>
<td></td>
<td></td>
<td></td>
<td>58.85</td>
<td>59.11</td>
<td>58.57</td>
</tr>
<tr>
<td>Bangladesh</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barbados</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 2. Visualisation of the Gross Enrolment Ratio (GER) and Gender Parity Index (GPI) Datasets, Global Education Digest 2010, UNESCO Institute for Statistics.*
But wikis, blogs and other Web 2.0 technologies are much more than simply tools to facilitate collective contribution. They are the force behind a new era of mass collaboration and participation that is changing the way we deal with global issues. Information is now being created, updated, discussed and made public by an engaged community from all over the world. Rather than an expert or team of experts assembling information on a particular topic, access is now open to worldwide resources in developing knowledge around a subject. A recently coined term, “wikinomics”, defines this movement as “the new force that is bringing people together on the net to create a giant brain”. Openness and collaboration using Web 2.0 functionality provides for a democratisation of innovation. The speed, scope and manner in which innovation occurs has radically and rapidly changed since Web 2.0 tools, including wikis, can provide “low-cost collaborative infrastructures... in ways only large corporations could in the past”. Hence, the collective knowledge and ingenuity of individuals and businesses is fully exploitable.

3. What are Wikiprogress and Wikigender?

Wikiprogress (www.wikiprogress.org) and Wikigender (www.wikigender.org) are interactive communication tools that centralise information, data, initiatives, publications, events, media coverage and research networks that are all part of the international movement to look beyond GDP in measuring the progress of societies. Wikiprogress brings together many dimensions of progress (ecosystems condition, human well-being, economy, governance and culture). Wikigender specifically looks at gender equality and women’s empowerment, with a particular focus on developing countries.

3.1 Wikiprogress

Wikiprogress is the official online platform for the OECD-hosted Global Project on Measuring the Progress of Societies. It was announced and presented in beta version at the 2009 OECD World Forum in Busan, Korea. Since then, Wikiprogress has grown significantly: it now has over 450 active members, over 840 articles and over 8,400 unique visitors a month. The Global Project is a partnership project that aims to develop a large "network of networks" of experts in the field of research into progress, including academics, civil society and NGOs, to work together on gathering information and data related to measuring progress. The wiki platform facilitates this collaboration and participation, ensuring the movement is inclusive and ever-growing.

It is not only the community that makes Wikiprogress effective, but also the content diversity. So far, the term "information" has been used to cover all of the content on Wikiprogress, but to understand the extent of the information available, it can be divided into key sections. These divisions will change and develop over time as the Wikiprogress community expands and enriches the content.

- **Information by topic** – Information on a particular dimension of progress (e.g. peace, trust, pollution) is gathered in one article with internal links to related pages, external links to relevant organisations and a list of reports and publications for further reading.
- **Information by country** – Each country has its own article listing the main national and regional initiatives. The country articles also link to relevant data and list country-specific publications and reports.
- **Progress initiatives** – Initiatives from around the world that gather information and/or data on measuring progress have their own article including the aim, function and background of the
particular group. These articles also contain links to relevant data, videos of presentations and speeches, publications and reports.

There are many different initiatives, some working within specific dimensions of progress and others working across all dimensions.

- **Progress publications** – Wikiprogress highlights reports and papers as they are published, on the condition that they are not for sale. The reports can be searched by topic or by country.

- **Progress-related events** – Events are held worldwide focusing on different dimensions of progress, the measurement of progress and the development of better indicators.

- **Media coverage** – A good measure of a movement’s success, and certainly a good way to monitor the development of the movement, is by tracking what the media is covering and examining media treatment of key players and events. The Wikiprogress Community Portal gathers news items and blog posts on the movement as they are released.

- **Finally, one of the most significant aspects of Wikiprogress:**

  progress-related data. This component of Wikiprogress is detailed in the last part of this paper which focuses on data.

### 3.2 Wikigender

Wikigender is an online platform that was developed by the OECD Development Centre and launched on 8 March 2008, on the occasion of International Women’s Day. It is the first wiki platform hosted by the OECD, and acted as a pilot project before the launch of Wikiprogress. Wikigender is therefore naturally linked to Wikiprogress. Furthermore, measuring the progress of societies without paying particular attention to women, who represent more than half of the population, would not be offering a comprehensive picture.

Since its launch, Wikigender has seen its community grow exponentially: it now consists of over 960 registered users coming from around 170 countries and comprises more than 1,120 articles. Added to these statistics, the site recently attracted over 15,000 visits during the course of a single month, a number that increases monthly.

Wikigender is an interactive online tool that focuses on gender equality issues, as well as on data and measurements in the area of gender equality. Like Wikiprogress, the site allows its registered users to add, edit or discuss information and/or data provided on the site. The platform therefore easily allows for up-to-date and accurate information.

The information available on the site is also organised by topic/category, country, organisations/initiatives, publications, events and statistics. It also has a community portal that lists media coverage on timely gender equality issues.

In terms of specific projects, Wikigender has started to work with Paris-based universities, engaging students to add content to the site and benefiting from the networking opportunities of Wikigender. This programme, called “Wikigender University”, will be extended to selected universities worldwide, further enriching Wikigender with content related to the measurement of progress made in gender equality, and at the same time empowering students through information-sharing and capacity-building in the area of information technology.

Similarly, “Wikigender Impact” is a project under development examining the establishment of a bridge between the community of donors and policymakers and the community of practitioners on the ground, by developing a database of case studies and becoming a clearing-house for funders and those seeking financial support for projects.

### 3.3 The importance of understanding data

Developing better indicators and awareness is not enough. We also need to ensure that the measures become widely used, and widely understood – not only by statisticians, but by all those interested in societal progress.

Visualising data in a time series allows us to see the impact of the progress – or regress – according to the particular data set. It also makes the data easier to understand, so that a wider audience has access to understanding its significance. There follows an introduction to the data presentation and visualisation tools available on the wikis.

#### a) Wikiprogress.Stat

It is one thing to have good statistics. It is another to make sure these statistics are used. New Information Communication Technology (ICT) tools have the ability to not only make data easily available for anyone with an internet connection, but also allow for the data to be presented in a way that can be understood. Both Wikiprogress and Wikigender have developed ways in which data could be better understood.

Wikiprogress has developed a statistical tool that facilitates the task of data collection, analysis and dissemination: Wikiprogress.Stat. Data from a variety of sources is progressively populating this database of progress indicators and is being used to support articles in both Wikiprogress and Wikigender. The database also complies with the wiki model; users can upload their own data and metadata, therefore democratising the process of data collection. Through Wikiprogress.Stat, anyone can upload data and access the information. After a quality control procedure, the data will be uploaded into Wikiprogress.

The overall aim of Wikiprogress.Stat is to create a robust database of progress indicators for a wide range of regions, over the longest possible time-scale. There are currently over 100 data sets in the Wikiprogress.Stat database, including The Global Peace Index, World Bank Economic Indicators, and The World Database of Happiness, amongst others.

Tables, graphs and charts created on Wikiprogress.Stat will then be used in articles on Wikiprogress to illustrate how different progress dimensions have impacted the well-being of particular societies.

Additionally, the OECD is working with Norrköping Communicative Visual Analytics (NComVA) on the implementation of a new data visualisation tool, eXplorer, to allow users to create stories using data sets in Wikiprogress.Stat and visualise the material in an animated time series. Each visualised story, a vislet, is embeddable in Wikiprogress articles and on external websites, blogs, etc.

Furthermore, the eXplorer will be an embeddable tool not only in wiki articles, but also in external blogs and online journals. This will broaden the scope of discussions and reports that can use the visualisation to assist in explaining and/or supporting ideas and arguments on social progress.

#### b) Wikigender

Gender equality datasets are also uploaded through Wikiprogress.Stat. Apart from presenting OECD data such as the Gender, Institutions and Development Database (GID-DB) or the Social Institutions and Gender Index (SIGI), Wikigender links to www.my.genderindex.org, a tool allowing users to manipulate SIGI data for their own analyses.

Furthermore, any new gender-related data included in Wikiprogress.Stat automatically appears on Wikigender. This is the case, for example, of UNESCO data accompanying the 2010 Global Education Digest.

Finally, Wikigender also presents information on the gender equality situation by country. Each country page includes quantitative and qualitative data from the SIGI as well as data from additional sources to offer a more comprehensive view of gender equality in a given country. For example, SIGI data examines discrimination against women from the point of view of social norms, cultural values, and traditions. But some country notes also cover data from Wikigender partners: the World Bank and the International Finance Corporation (IFC), looking
at laws and regulations affecting women’s prospects as entrepreneurs and employees. More data will also soon be added, for example, examining women’s and men’s access to land. Wikigender will also continue to collaborate with other organisations in order to give the most accurate and comprehensive picture of gender equality across countries. Such an approach gives the advantage to the user to find information on one topic, in one country, looked at from various angles and perspectives.

The above tools offer various entry points to the data, allowing any site users to access, upload, manipulate, or analyse data and draw analytical conclusions. Both Wikiprogress and Wikigender have large and diverse audiences which include policymakers, non-governmental organisations, think tanks, international organisations, statistical offices, students, academics, practitioners, and donors.

Conclusion
The movement to look beyond economic indicators in measuring well-being can be successful by working society-wide. In order to measure what is important to societies, a collaborative effort is required. Currently, wikis are the correct web platform for collective brainstorming, collaborative problem-solving and for centralised global efforts and initiatives working towards a common goal. In this regard, Wikiprogress is the appropriate tool for the successful collection, analysis and dissemination of information across countries and societies, as Wikigender similarly for collaboration and knowledge sharing in the area of gender equality and gender statistics.

We invite members of the IASSIST community to join the community and participate in the ongoing debates on www.wikiprogress.org and www.wikigender.org. We are always looking for new ways to collaborate with different organisations. If you are interested in either of these initiatives, we invite you to contact us at info@wikiprogress.org.

Notes
5. Ibid.
Secure Data Service:

An improved access to disclosive data by Reza Afkhami, Melanie Wright, Mus Ahmet

Abstract
The Secure Data Service is a secure environment funded by the ESRC to provide researcher access to disclosive micro data either from their offices, safe rooms in their institutions or on site at the UK Data Archive. Operation is legally framed by the 2007 statistics Act which makes possible access to the confidential data for statistical purposes. This short paper introduces this new UK Data Archive service and proposed specifications, as well as challenges facing data service providers. We envision that the proposed SDS infrastructure will meet the requirements of the data security model. The paper also aims to be an exemplar for a secure remote access practice.

Keywords: Remote access, Data security, Citrix technology, Secure Data Service

1. Introduction
Disclosure of personal information can be harmful and may result in denial of services, embarrassment and loss of reputation and trust which in turn reduces the response rate and jeopardises the future research. Research results based on disclosive data can also cause indirect harm by affecting perceptions about a group to which a person belongs.

Balancing data confidentiality and legitimate requirements of data users is a key problem of the Secure Data Service (SDS). Confidentiality of individual information can be protected by restricting the amount of information provided by adjusting the released Microdata /tables/ statistical outputs (restricted data), by imposing conditions on access to the data products (restricting access), or by some combination of these.

The UK Data Archive Secure Data Service is a new service to allow controlled restricted access procedures for making more detailed Micro data files available to some users (Approved/Accredited Researchers), subject to conditions of eligibility, purpose of use, security procedures, and other factors associated with access to the SDS data.

Building on the success of other secure data enclaves worldwide, and employing security technologies used by the military and banking sectors, the SDS will allow trained researchers to remotely access data held securely on central SDS servers at the UK Data Archive. The aim of the service is to provide approved academics unprecedented access to valuable data for research from their home institutions, with all of the necessary safeguards to ensure that data are held, accessed and handled securely.

The SDS follows a model which recommends that safe use of data should include safe project, safe people, safe setting and safe output (Ritchie, 2006. see Figure 1). To achieve this goal, data security depends on a matrix of technical, legal, contractual, and educational factors. The structure of this paper revolves around these factors, demonstrating how SDS has set up the necessary infrastructure to meet these requirements.

We discuss the legal and contractual responsibilities of the users and their institutions followed by issues such as user education and training prior to data access. The technical features and the system specifications are also examined. We also discuss the kind of disclosive data we aim to support in the SDS. Finally, the challenges facing the SDS operation will be examined.

2. Legal and contractual framework
Users of the SDS will be required to be either “ONS Approved Researchers” or “ESRC Accredited Researchers.” The first of these is defined by the Statistics and Registration Services Act 2007 as “an individual to whom the Board has granted access, for the purposes of statistical research, to personal information held by it.” No definition currently exists of an “ESRC Accredited Researcher,” but we assume that it will have a similar status to an ONS Approved Researcher. This is a person who has been granted access for the purposes of statistical research to personal information which has been licensed to the ESDS/UK Data Archive/University of Essex for dissemination on behalf of a government department or some other data provider. Neither of these two types of user will be able to use the SDS without appropriate training. Mandatory training will allow the UK Data Archive to ensure that end-users are fully aware of any penalties which they might incur if they cause a breach. We believe that if there is user approval to any penalties for breaches, and that they believe that these penalties are reasonable and necessary, we will avoid the inadvertent disclosure to which social science researchers are most likely to be prone.

The 2007 Act also allows for increased sharing of data between ONS and other Departments, subject to agreement by Parliament on a case-by-case basis. At the same
time the Act also outlines measures designed to protect the confidentiality of personal information. The Act states that a person who discloses personal information "is guilty of an offence and liable — (a) on conviction on indictment, to imprisonment for a term not exceeding two years, or to a fine, or both, (b) on summary conviction, to imprisonment for a term not exceeding twelve months, or to a fine not exceeding the statutory maximum, or both."

The SDS will immediately suspend access to the service if it believes that any user is perpetrating or attempting to perpetrate any of the breaches listed in SDS security breaches or SDS confidentiality agreement. A full investigation will follow.

Users will be required to complete an on-line form which collects personal and institutional details, information about their proposed data usage, and information which demonstrates their expertise and ability to conduct the research described in a competent and secure manner. They will also be required, if they have not already done so, to agree and sign the standard UK Data Archive End User License (EUL), and also agree/sign any Special License conditions which apply to the resource they wish to access. This application will be first checked for accuracy, sense and completeness by UK Data Archive staff, and then forwarded to the data owners for their access authorisation.

Once authorised, users would be informed and requested to sign up for appropriate training (if they have not already been trained). Upon completion of training, users would be granted permission to access the secure data server, either from their own desktop if the owners of the data they wish to use permit, or from their institution’s secure data access room. If their institution does not have a secure data access room, the user will have the option of negotiating access from another nearby institution’s safe room (the SDS will offer ‘matchmaking’ introductions, but the specific arrangements must be under the control of the institution holding the room, as audit trail security will be their responsibility) or coming to the University of Essex, to access the service onsite.

3. Education & Training

Researchers are the known weakest links in data security. Education coupled with the stricter legislative protections mentioned above, can offer another potentially efficient means of improving confidentiality, as disclosure probability can be decreased without imposing costs on rule-abiding researchers.

Before becoming an active user of the SDS, users will have to attend a mandatory training session which will focus first on the user’s legal and ethical responsibilities within their SDS user license agreement, the mechanics of how to use the SDS, what they can and cannot do in a remote access setting, and the potential of the collaborative spaces. The second part will focus on principles of the statistical disclosure control, assessment of outputs, and analysis aspects of the particular datasets in the SDS.

Access to the SDS will only be granted after users have attended an SDS training session. SDS staff will vet data analysis outputs for disclosure issues, to ensure that nothing escapes the secure data setting, which could compromise the data security (safe output). One of the purposes of the training is to give researchers the ability to recognise confidential data and distinguish it from statistical results that are safe to remove from the SDS. In effect, the training removes the ‘reasonable belief’ defence for a disclosure. We believe that penalties will only be an effective deterrent if they are known, and it should also be clear that we are more concerned with prevention than punishment.

4. System Specifications

The technology used by the SDS must be secure and the system adheres to the highest standards of quality. The technical model that has emerged is one which shares many similarities with both the ONS VML and the NORC Secure Data Enclave. It is based around a Citrix infrastructure which turns the end user’s computer into a remote terminal giving access to data, statistical software, and collaborative spaces on a central secure server held within the UK Data Archive. The system is flexible, in that depending upon the wishes of the data custodians, access can be restricted to particular users (safe people) and/or particular locations (safe rooms/machines). It is secure because all data manipulation occurs on the server, which is maintained to very strict security protocols.

Beyond the general security policy, the secure server itself will be subject to additional security measures and controls. Approved researchers will access the proposed SDS by using VPN (Virtual Private Network/thin-client) technology, which encrypts the data transmitted between the researcher’s computer and the host network. Other components of the VPN technology allow control to be established over which network resources the external researcher can access on the host network. The service will employ a Citrix XenApp server farm, which participates on two networks Mulcahy et al, 2008).

How the system operates?

With this technology, although all applications (SPSS, STATA, etc) and data run on a central server at the UK Data Archive/SDS, the Approved Researcher still interacts with a full Windows graphical user interface. This means that the researcher never has to install any complex applications on his/her remote computer, the only application required by the Approved Researcher is a web browser. This also means that the UK Data Archive can prevent the researcher from transferring any data from the data archive to a local computer. For example, Citrix can be configured so that data files cannot be downloaded from the remote server to the user’s local PC. Similarly, the Approved Researcher cannot use the ‘cut and paste’ feature in Windows to move data from the Citrix session into an Excel spreadsheet sitting on the local computer. Finally, the user is prevented from printing data from a local computer. The Approved Researcher logs onto the SDS system remotely via a web secure (HTTPS) browser. All data processing is carried out on a central secure server, which processes all requests centrally and returns information about the results. No data travels over the network, except the statistical results sent from the central server to the remote location by an encrypted email after the final outputs are checked against statistical disclosure controls.
Key Features
- Clients cannot remove data
- Absolutely no webpage access
- Clients cannot import data
- Data transfers are logged
- All traffic is encrypted
- Smart Auditing
- Critical Security updates are applied daily

5. Benefits
This system may pose an inconvenience to the user compared to their accustomed ability to use EUL (End User License- similar to Public Use File) data on their desktop with all their favourite local software and networked resources. However, it is a price users will gladly pay for local access to data which they might otherwise have had to travel to ONS sites to access, or simply have been unable to access. The SDS will benefit users in:

- A self-contained secure 'home away from home' service with familiar analytical environments;
- Ability to work in their own private work areas or in shared areas with other approved researchers;
- Access to enhanced, highly sensitive available data storage in tandem with the related metadata through increased capacity and environmental protection;
- Possibility of data linkage exercise with using existing data in the UK Data Archive or other administrative data subject to approval of data owners/custodians; and
- Collaborative functionality including survey and document library, SPSS/ STATA code library, knowledge repository, disclosure review and technical assistance.

6. Data
A variety of data may potentially be available to users within the SDS. We are in ongoing discussions with the owners of key sensitive data resources about how SDS might assist in broadening the use and utility of these important resources, whilst assuring that legal, moral and security requirements are met. The specifications of data candidates include:

- More detailed variables from existing ESRC-funded data resources;
- More previously unavailable detailed variables from government social surveys;
- Other government data previously only available in onsite enclaves, or previously unavailable to academic researchers altogether;
- Business data which has commercial sensitivity;
- Administrative data; the SDS may be able to provide a secure environment for data linkage activities to researchers whose home institutions lack the technological wherewithal to offer it (restrictions must be negotiated and approved by the data owners/custodians upon user's application); and
- Data previously considered too sensitive or potentially disclosure due to its very nature, such as longitudinal data, medical data, etc.

In addition, the service will allow users to bring in less disclosive data from the UK Data Archive standard EUL holdings, upon request or researcher's own data subject to the standard ingest checks and approval.

7. Challenges
The two main goals of the Secure Data Service are: maximizing utility of microdata for research purposes, and protecting the confidentiality of individual respondents. Access to confidential data is an exception to the non-disclosure rule that must be justified according to the balance of the public good of the research against the risk of a breach of respondent privacy. Thus, the SDS hopes to maximize data utility while minimizing the disclosure risk, utilizing a strategy that is simple, well-communicated and acceptable to users.

However this task may be daunting as there is no consensus on the definition of what is safe data and second, even more contentious is what information loss means and how it can be measured. As any effort to implement confidentiality protection is associated with some loss of information.

Maintenance of confidentiality needs a consistent and coherent approach and we must trust the researchers as no environment is free of risk. For example, how can we prevent against a manual data copying or using photographs for researchers who remotely have access to the disclosive data or even user's memory. For the majority of researchers, data breach happens for access convenience and not out of a malicious intention and surely remote desktop access to the data would diminish that temptation. However, the possibility of disclosure is always there, the legal framework and training and education may deter users -who are approved researchers afterall - to perpetrate any confidentiality breaches.

8. Evaluation and monitoring of the outputs
Careful user vetting and the most secure analysis environment in the world cannot on its own ensure that data are not disclosed. The missing piece of the data security puzzle is not what goes into the secure data system, but what comes out of it. For the service to be able to meet the security guarantees placed upon it by the data guardians, it must offer some form of output screening. If an output has been determined to be disclosive, it will be up to the user to determine the best way to render it safe.

SDS adheres to European-wide ESSNet standards on good practice in statistical disclosure control of tabular and other statistical analytical outputs (Hundepool, et al 2009). The SDS disclosure advisor will divide outputs from SDS into two main categories:

- Safe: very low risk of disclosure – output will be released promptly
- Unsafe: high risk of disclosure – output will be blocked in its current form and won't be released, the researcher must produce safe outputs and demonstrate that they are free from the disclosure risks

There are several solutions available to protect the information of the sensitive cells:
• Combining categories of the spanning variables (table redesign). Larger cells tend to protect the information about the individual contributors better.
• Suppression of additional (secondary) cells to prevent the recalculation of the sensitive (primary) cells

9. Summary
The SDS is a secure environment funded by ESRC to provide researcher access to disclosive micro data either from their offices, safe rooms in their institutions or on site at the UKDA. It has two goals: to promote researcher access to sensitive micro data and to protect confidentiality. SDS operation is legally framed by the 2007 Statistics Act, which makes access to confidential data for statistical purposes possible.

Researcher access to microdata serves the public good both by leveraging existing public investments in data collection, and by ensuring high quality science through the replication of scientific analysis. The SDS provides Approved/Accredited researchers with remote access to microdata using the most secure methods to protect confidentiality. This is achieved by implementing technological security (Citrix gateway), applying statistical protections, enforcing legal requirements, and training researchers. The SDS also ensures that valuable data are preserved for the long term by documenting the data using DDI compliant metadata standards. In addition, the SDS aims to engage the research community in using its shared data space to share information which enables collaboration among geographically dispersed researchers.

References

Notes
1. This paper was presented at the IASSIST 2010 in the session "Secure remote access to restricted data". Corresponding author: rafkhami@essex.ac.uk, Dr. Reza Afkhami, Senior Data and Support Services Officer, Secure Data Service, UK Data Archive, University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, Tel: +44 (0) 1206 874968, Fax: +44 (0) 1206 872003
2. Secure remote access is also developing in Denmark, Netherlands and Sweden
A user-driven and flexible procedure for data linking

by Cees van der Eijk and Eliyahu V. Sapir

Abstract
Over the past decades, social scientists enjoyed a rapid increase in the availability of various types of data. Many of these pertain to different aspects of the same multifaceted phenomenon. In the field of electoral studies, for example, there are data about citizens, political elites, party manifestos, media and the context within which elections take place. Researching such multifaceted phenomena requires this diverse information to be analysed jointly, rather than separately. That, in turn, requires the linking of separate datasets (also known as data fusion, or conflation), which is largely a relational database (RDB) management problem. In spite of its large and rewarding potential for empirical research, data-linking is practiced relatively little in the social sciences. This is largely caused by lack of relevant training amongst social scientists to cope with the methodological and technical difficulties involved in data linking. This article presents an approach for facilitating data linking without requiring additional training of researchers. It defines necessary RDB operations as a structured series of user-choices, to be included in a user interface that generates and implements the RDB operations once all choices have been made. The advantage of this approach over the public dissemination of integrated datasets constructed by ‘experts’ is that it does not assume that ‘one-size-fits-all’; it is flexible and tailored to the needs of end-users. This approach can be applied in a wide variety of contexts. An implementation is under development for the PIREDEU project in the field of comparative electoral research.

Keywords: data linking, data integration, relational database, PIREDEU, electoral research.

Introduction: An embarrassment of riches
Compared to only a few decades ago, comparative social researchers now enjoy the availability of a wealth of datasets. If they are interested in behaviours, attitudes and orientations of citizens, they will find that in addition to sundry ad hoc surveys, many regularly conducted national election studies, general social surveys, household, labour-market, crime, and other surveys are available in an increasing number of countries. Moreover, they will find an ever growing number of explicitly comparative surveys that are repeatedly conducted in multiple countries, thus enabling comparisons across national contexts as well as over time. These include, amongst many others, the Comparative Study of Electoral Systems (CSES, covering up to 38 countries), the World Value Studies (WVS, up to 87 countries) and European Value Studies (EVS, up to 45 countries), the International Studies of Political Psychology (ISPP, up to 45 countries), the European Social Surveys (ESS, up to 31 countries), and the European Election Studies (EES, up to 27 countries).

This wealth of empirical material is not restricted to any particular social discipline, nor is it only made up of mass surveys. For reasons that will become apparent later in this paper, we are particularly interested in studies relating to elections, parties and public opinion, but our observations of that domain can easily be generalised to other large areas of social scientific inquiry. When reviewing our own field of interest, we find, increasingly, that in addition to the many available mass surveys, political and social elites of various kinds are surveyed as well, in single countries or comparatively across a number of countries. Beyond the domain of surveys, we find data derived from party manifestos covering almost all parties that have ever competed in democratic general elections after World War II. These have more recently been complemented by the Euro-manifesto program that codes the contents of the manifestos of all parties that have ever competed in direct elections of the European Parliament (EP). Content analyses are also increasingly used to generate systematic data about media communications, and include projects...
that yield data comparable over time and across countries. Other extensive data sets have become available for yet other organisations and institutions, such as social movements and pressure groups. At the level of states, there is also an abundance of data pertaining to sundry economic indicators, political and social indicators, formal institutional arrangements, government performance, and so forth.

Each of these data collections by itself provides rich possibilities for empirical research, and indeed we see increasing numbers of publications making use of this potential. Yet the promise that these data hold is much greater if they can be linked to each other. In recent years, many of the principal investigators of these large data collection efforts have referred to this larger potential as part of the justification for investing in these costly enterprises.

Indeed, many of the most interesting questions in social research do not pertain only to citizens, or only to elites, or only to media, and so forth. Rather, they have to do with the interactions between various kinds of actors, organisations and institutions, which are affected by the characteristics of different contexts, or with the social, political and economic consequences of these contextualised interactions. As a case in point, questions about the quality and functioning of representative democracy pertain simultaneously to citizens, political parties, political elites, and mass media, among other things. Seemingly simple concepts such as accountability and representation relate to the interaction between citizens and (political) elites, as well as various types of processes that involve the media (for example, the effects on citizens and elites of agenda-setting, framing, priming, and spin and hype). To the extent that the functioning of representative democracy is affected by economic developments, all of these interactions and relations have to be contextualised in economic terms. From a dynamic perspective this leads to questions as to whether ‘the economy’ is an autonomous factor affecting the behaviours and interactions of the various actors in democratic processes, or whether it is endogenous and the consequence of these behaviours and interactions.

Important questions that require empirical information from a variety of different actors, groups, organisations, institutions and contexts exist in all social sciences. They may focus on, for example, social integration, crime, traffic and mobility, or the efficiency of markets, but they all have in common that they cannot be adequately addressed using information pertaining to only one of the interacting actors and institutions.

Our ability to address important multi-facetted questions has not only been increased by the availability of abundant relevant data, but also by advances in multivariate analytical methods, software and affordable computing power. Complex models that until recently were beyond the computing infrastructure available to most researchers can currently be estimated on standard personal computers using generally available software. Of particular importance for empirical social research are the advances in ordinal- and nominal-level multivariate analysis, latent structure modelling, structural equation modelling, dynamic modelling and multi-level methods.

In view of the wealth of relevant data and the availability of tools to analyse them, one might expect current social science literature to abound with publications that join together information from multiple data sources in order to more effectively address the important and broad-ranging questions referred to previously. Yet, such publications are fewer than one would expect, which creates a somewhat puzzling and embarrassing situation.

**Diagnosis**

In principle, many separate datasets can be linked in ways that would allow important research questions to be addressed in more powerful ways than analysing each of these resources separately. If these separate datasets relate (by way of their units or their variables) to the same real-world objects, such as countries, political parties, media outlets, and the like, they can be seen as component parts of relational databases (RDBs). The methodology of RDBs is well developed and provides a multitude of ways to generate joint information from different components that provides a richer base for analyses than the sum of the separate parts. Moreover, RDB software is widely available. Why, then, do we see so very few efforts of integrating or linking different datasets? A number of factors contribute to this state of affairs, including the following (without claiming to be exhaustive):

**a) Lack of harmonization.** Linking of separate datasets in a RDB requires the same objects being identified in the same way in each of them. Frequently this is not the case. As a case in point, identification codes of political parties differ more often than not between successive editions of a series of national election studies, each of which pertains to a different election. Even data infrastructures that pride themselves on their over-time comparability of coding often fall short in this respect. This problem is not limited to the identification of parties, but also to countries, regions, media outlets, and so on. As a consequence, any linking has to be preceded by a complex and costly data harmonization stage. Without dedicated resources it is impossible for most researchers to undertake such projects, and funding agencies see little glory in providing grants to produce such ‘continuity’ datasets. The few that do exist are not extended when new studies are released and become outdated, thus losing their relevance. Analysts that do aspire to the simultaneous use of data from different sources thus have to make their own tailored ‘solution’, which is often too narrowly focused to suit the needs of others. It is therefore not surprising that, when faced with such obstacles and with publication requirements from their own universities, many opt for the short-term option to analyse a single dataset and forego the potential riches that could be gained in the long term from data-linking.

**b) Limitations of ‘statpacks’ and other statistical software.** Much of the statistical software used in the social sciences is bundled in so-called statpacks such as SPSS, SAS, STATA and R. These packages contain a wealth of statistical procedures, as well as extensive procedures for data management, such as recoding and creation of new variables. Yet, they are fundamentally geared towards the analysis of ‘flat’ rectangular data matrices, and their capabilities for managing RDB information range from non-existent to extremely limited and cumbersome. As a consequence, after separate databases have been harmonised and linked in a RDB structure, dedicated RDB management tools must be used to generate the (rectangular) data matrices that lend themselves to analysis with the analytical software social researchers have been trained to use. This not only requires additional work, but requires working with software that is often unfamiliar to many researchers in the social sciences.

**c) Lacunae in social science research training.** Research training in the empirical social sciences traditionally focuses on questions of general research design, data collection methods, and multivariate statistical data analysis with statpacks and similar software. Many researchers, therefore, are well-versed in advanced multivariate modelling procedures, yet woefully untrained in recognising the potential benefits of RDB in managing data productively. When confronted with multiple datasets, it seems that many otherwise excellent researchers see only a collection of separate datasets, each of which can be analysed in sophisticated ways, but individually. What they often do not see are the ways in which separate datasets...
can be linked in a RDB, which, in turn, can be used to generate new rectangular data matrices that provide better platforms for addressing the substantive research questions they wish to pursue.

None of these three obstacles to linking and merging data from different sources is insurmountable, yet they are not easily overcome as they are rooted in entrenched traditions of training and acquired routines. Unleashing the full potential of linkable data thus requires more than just pointing out the benefits to be gained. It also requires infrastructure, in the form of software that offsets the lack of RDB familiarity amongst social scientists and that can be used without extensive further training.

This paper describes an attempt to work around the obstacles that often prevent analysts from linking data from various sources. Although this attempt is, in principle, not limited to any particular kind of research problem, or to any particular collection of datasets, we nevertheless present it in the context of its development, the PIREDEU program. Moreover, as our work is still ongoing, our presentation in this paper relates to ‘work in progress’, with some parts having being developed already, and others still under development (see also van der Eijk and Sapir 2010).

PIREDEU

PIREDEU is a program of research funded by the European Union (EU) under the Seventh Framework Programme from 2008 to 2011. This three-year design study assesses the feasibility of upgrading the existing European Election Studies to a research infrastructure for studies into citizenship, political participation, and electoral democracy in the European Union. The scientific and technical feasibility of this infrastructure is elaborated by means of a pilot study conducted in the context of the 2009 elections to the European Parliament.1 In contrast to many other comparative research programs about elections that only investigate voters, or only parties, PIREDEU considers its subject matter, electoral democracy, as a complex set of interactions between voters, parties, candidates, media and relevant institutions (such as election rules). It therefore collects empirical information concerning different units and it does so for each one in the most appropriate manner. The various data components of PIREDEU are:

• A voter study that conducts surveys of representative samples of approximately 1,000 respondents each from the electorates of the 27 member states of the EU.2 Apart from translation and reference to country-specific institutions, the questionnaires were the same for all countries. The questionnaires covered three main themes: 1) electoral behaviour and party preferences; 2) political attitudes and orientations; and 3) background characteristics and media usage.

• A candidate study that conducts surveys of the candidates listed on the ballots of the European Parliament elections of 2009 in each of the 27 member states of the EU.3 With the same provisos as mentioned for the voter survey, the questionnaires were identical for the candidates from each country and from each party.

• A manifesto study that consists of coding the content of the election manifestos of the political parties vying for votes in the 2009 European Parliament elections.4 The coding units are sentences or quasi-sentences, each of which is classified in one of many content categories relating to a wide range of policy domains: external relations, freedom and democracy, political system, economy, welfare and quality of life, fabric of society and social groups, and European integration (cf. Wüst and Volkens 2003). After aggregation to the level of parties, this provides data on the relative emphasis that parties place on the substantive topics reflected in the coding categories.

• A media study that consists of coding the content of media items during the three weeks leading up to the European Parliament election.5 All news items were coded for the most important TV news programs and the most important newspapers in each of the 27 EU member states. Each news item was coded on a large range of characteristics, including topics, actors displayed, use of frames, and physical characteristics (length, placement, embellishments, etc.).

• A contextual information study that brings together information at the level of the 27 member states of the EU, regarding the results of the European Parliament election and other recent elections, electoral procedures, voting rules and other relevant institutions, the incumbent government, and economic conditions.6

In line with PIREDEU’s perspective on electoral democracy, these various datasets are seen as providing complementary information about a complex and multifaceted reality. And although it is perfectly feasible to analyse each of the datasets in isolation, one of the main tasks of the program is to link or integrate them in user-friendly ways, thereby promoting a fuller utilization of the joint potential of the data, and, ultimately, better research on electoral democracy in Europe.

In order to achieve this objective, great care was given to assuring that the variables by which the data from the different components can be linked – the ‘keys’ in RDB terminology – were coded in exactly the same way in each component. This relates in particular to the identification of countries, political parties and media outlets. All components are structured by country and all relate to political parties. The voter study includes questions about party choice, party preference and perception of parties; the candidate study explores the relationship between candidates and the party for which they are listed on the ballot, and other questions about their own and other parties; the

![Figure 1. The PIREDEU Relational Data Base](image-url)

**Figure 1. The PIREDEU Relational Data Base**

In order to achieve this objective, great care was given to assuring that the variables by which the data from the different components can be linked – the ‘keys’ in RDB terminology – were coded in exactly the same way in each component. This relates in particular to the identification of countries, political parties and media outlets. All components are structured by country and all relate to political parties. The voter study includes questions about party choice, party preference and perception of parties; the candidate study explores the relationship between candidates and the party for which they are listed on the ballot, and other questions about their own and other parties; the
manifesto study explores the relationship between each manifesto and its author (i.e., party); the media study examines coded news items in terms of parties being mentioned and evaluated; and, lastly, the contextual data include the identification of the party/parties that form the incumbent government, and the results of various elections in each of the countries. The identity of media outlets is crucial in the media study to identify the outlet from which each coded item was taken, and in the voter study to identify media outlets that respondents use for their information. These keys define the primary relationships in the PIREDEU RDB, as illustrated in Figure 1.

The solution that we chose for allowing data from the various components to be linked in a rectangular data matrix that can be analysed with the statistical software mostly used in social science research will be described below, but to clarify the task at hand we first present an example of a substantive research question that requires information from all components to be merged into a single analysable file.

A substantive example of the need for linking data from PIREDEU components

Suppose we are interested in the orientations and behaviour of candidates, and, more particularly, in how salient various issues are for candidates. This information is specific for the candidates, and available from the candidate survey.

Were we to analyse issue salience for candidates only from the data in the candidate survey, we would model the variance in the dependent variable (salience of issues for candidates) in a multi-level model with candidates nested within parties, which, in turn, are nested within countries. As independent variables at the candidate level we can use all other variables collected in the candidate survey, including various political orientations and background characteristics, and the identity of the party and the country of each candidate. In the absence of further information about the traits of these parties and countries, their impact would be modelled in a random effects specification.

Such a random effects model would be unsatisfactory because it would only tell us that some of the variance in issue salience at the level of individual candidates can be attributed to parties and to countries, but we would be in the dark as to the form of this relationship. That is, which kinds of parties and which kinds of countries have a positive or negative impact on issue salience?

The model could be made more informative by adding information about parties and about countries, thus allowing a mixed model specification in which the explicature characteristics of parties and countries are modelled as fixed effects and the remaining variance at that level as random effects. This additional information can be derived from the other data components of PIREDEU.19 In the absence of any infrastructure or specific tools, such information would have to be added manually by the analyst. From the manifesto study one could, for example, derive how much emphasis parties place on each of the issues in their manifestos. This information would be added to the candidate dataset via a tedious procedure involving a large number of conditional statements. With some 200 political parties this procedure would be prone to error, and would take several hours to accomplish even for an experienced data analyst. In a similar way, one can add country information to the candidate file (which would be somewhat less onerous as there are only 27 countries). Additional information can be added that originates from the voter study or from the media study. The work involved becomes increasingly more complex if the theories that we want to test involve a wider set of relationships between candidates, parties, voters, media and contexts. Consider the following elaboration, which, although used here for its illustrative value, would substantively be neither unrealistic nor excessively complex.

The dependent variable – the salience of various issues for candidates – can be seen as a function of:

- The difference between the candidates’ personal views on issues and those of his/her party as expressed in its manifesto.
- The salience of the various issues for various media outlets, moderated by the extent to which potential voters for the candidate’s party are exposed to those media outlets. This expectation could be based on the notion that, politically, candidates cannot afford to ignore issues that are played up in the media, particularly if their own potential voters are exposed to the contents of those media. To test this, one has first to arrive at a measure of issue salience for media outlets, which has to be derived by some form of aggregation from the news items that have been coded. This outlet-issue-salience then has to be added to the candidate survey using country as key, as candidates are not directly linked to media outlets. Subsequently, the voter survey has to be used to distinguish potential voters for each of the parties, and then to determine for each of these groups the extent to which they are exposed to media outlets. This information has to be added to the candidate dataset using media outlet and party as keys.
- The salience of various issues for voters, moderated by their propensity to vote for the party of the candidate in question. This would be based on the expectation of candidates being responsive to voters in general, and in particular to voters who are likely to vote for their party. To implement this, one would have to use the voter study to distinguish voters according to their propensity to vote for each of the parties, and then to assess how salient the various issues are to each of these groups. This aggregated information would then have to be added to the candidate survey using party and country as keys.
- The effects of the factors listed in the previous bullet points are potentially moderated by country-contextual factors, such as the (temporal) location of the EP election in the domestic electoral cycle (for theoretical foundations of this expectation see van der Eijk and Franklin, 1996). This would require retrieving the relevant information from the contextual information dataset, and then adding said information to the candidate dataset using country as key.

After performing all these data operations, the candidate dataset, extended with information from the datasets pertaining to manifestos, voters, media and contexts, can be used to perform the desired multi-level mixed effects regression with cross-level interactions. Again, without specific tools or infrastructure to help accomplish these tasks, the required data management would easily take days of error-prone and tedious work. Moreover, all investments in that work would only be relevant for this particular question, and similar, but substantively different operations, would have to be performed for other substantive questions.

Managing the linking problem

In our view, any attempt to facilitate productive linking of data from different sources (or in our particular case, from different components of the PIREDEU program) has to recognise the following considerations and constraints:

- In terms of the outcome of the linking process – a data matrix that can be analysed by the kind of statistical software used by social scientists – there is no single or one-size-fits-all solution. What has
to be linked, and how exactly, is different for various substantive research questions. Any attempt to impose a single solution would be futile, as researchers will not use it if it does not fit their aims and theoretical and conceptual perspectives.

As a consequence, facilitating linking has to take the form of providing tools to accomplish the task, and not providing a linked dataset. This holds equally for linking data pertaining to different observation units, as for linking data pertaining to the same kind of observation units (e.g., repeated studies of the same kind).

- The outcome of the linking process must be a datafile that lends itself to analysis with the statpacks and other statistical software that is ubiquitous in social science research. As such statistical software is generally not capable of handling RDB structures, the linking process must result in a flat, rectangular data matrix.

- Many excellent empirical social science data analysts have not been trained in RDB management. Therefore, tools to facilitate data linking should not assume such familiarity and must provide, in a structured manner, the kinds of options available. As a consequence, relevant tools must disaggregate the linking process into successive tasks of limited complexity and clarify the options available at each for the user.

In accordance with these considerations, we chose to develop tools for linking data across the various PIREDEU data components in the form of a structured user interface that guides the user through a set of choices. The entire sequence of choices generates a syntax that specifies the required RDB operations and, when implemented, yields as an outcome the desired dataset with linked data. Actually, in view of the software considerations referred to above, it produces a tailored dataset into which information from other datasets is merged.

We will illustrate this approach by focusing on only two of the PIREDEU data components, the voter study (VS) and the manifesto study (MaS) (see Figure 1). Linking other combinations of data components operates along the same lines and will not be elaborated in this paper.

**Figure 2. Flowchart of user decisions involved in merging voter study data into manifesto study data — [MaS & VS]**

**Linking and merging the Voter Study (VS) and Manifesto Study (MaS)**

The units in the VS are individual respondents. The primary key in the VS is the respondent-ID. The units in the MaS are political parties, and the primary key in the MaS is the party-ID. The relationship between these two datasets is defined by a number of foreign keys in the VS that relate to survey questions about parties (each of these foreign keys is coded in the same way as the primary key in the MaS). These questions are about different matters, including actual party choice made in particular elections, attractiveness of parties as options to choose in a particular election, generalised affect, and perceptions of parties, among other things. What they all have in common is that their possible responses are defined in terms of parties. These foreign keys can be used for merging information from both datasets. In view of the considerations discussed in the previous section, this merging has to result in a flat data matrix that can be analysed by statistical software packages. This merging can therefore take two different forms, resulting in two different linked datasets: (1) a VS dataset (units are individual respondents) with information from the MaS merged into it, or (2) a MaS dataset (units are political parties) into which information from the VS is merged. Owing to the difference in the character of their units, these two forms cater to different research questions, and thus to different groups of researchers. Moreover, because the keyed link between the VS and the MaS is of a one-to-many type, the actual merging process is somewhat different when going from VS to MaS than the other way around.

**Merging MaS data into the VS.** The MaS offers information about the political parties that respondents mention in their answers to survey questions. As this information is not respondent-specific, it is identical for all respondents who refer to the same party in response to a question. Thus, for this information, one can regard the respondents as being nested, so to speak, in the parties they mention in their responses.

Merging is in this case a simple operation, as each mention of a party by a respondent relates to only a single case in the MaS data. When new variables are added to the VS, they contain the desired information from the MaS, linked by the correspondence between the chosen foreign key in the VS and the primary key in the MaS.

**Merging VS data into the MaS.** This kind of merging provides information about the composition of the groups of respondents who relate to the various parties in terms of choice, affect, particular perceptions, and so forth. This linked information may include anything available in the VS, such as respondents’ views on political issues, their social characteristics, media usage, political behaviour, and so on.

Merging in this case is somewhat more complex than in the previous case, as the relationship between each party and respondents generally involves multiple respondents. The variables to be added to the MaS file, therefore, have to contain summarizing information about the relevant group of cases in the VS. The analyst has to decide which of various possibilities is most desirable. Obviously, this is partly dependent on the measurement level of the relevant information in the VS. Means and variances would be relevant for interval level
variables, but that level of information is rarely available in survey data. For ordinal level data the ordinal equivalents of these summarizing parameters are available. At nominal level, summarisation is limited to proportions in all (or only in some) of the categories. The user interface for linking thus contains a set of choices for the analyst that specify the desired mode (or modes) of summarising VS data to be merged into the MaS.

Flow of end-user decisions. In order for the user linking interface to generate the desired tailored dataset, the analyst will be guided through a series of structured choices which are reflected in the flowchart in Figure 2. As a very first step, the user has to decide on the units that are to populate the desired merged file. In the example presented here, where we consider only the VS and the MaS, the question is whether we want to obtain a file of voters with merged data from party manifestos, or, alternatively, a file of parties and their manifesto data, into which voter information is merged. Figure 2 has been specified for the situation where the integrated file has parties as units, obviously a very similar, yet in detail, somewhat different flowchart would specify the decisions to be taken for the choice of individual respondents as units in the integrated file. Once the choice of units has been made, one should decide on the number of countries one would like to include in the integrated file. This number could range from 1 (i.e., a single country database as the final product) to 27 (i.e., all member states included). Next, the user should determine which parties will be included in the integrated file. This number is in the range of 1 (a single party dataset) to all parties participating in the EP elections (over 200). The next step is to define which respondents to include in the information to be merged (all respondents can be used to provide the information to be merged, or a selection that can be specified in terms of group criteria and weights). Then, the user should determine whether or not the variables to be merged should be recoded and, if so, how. Once these steps are completed, the user needs to define the aggregation parameters he/she would like to use in aggregating the VS data into the MaS data. In the user interface, the possible choices will be presented in the form of drop-down lists or of user-specifiable values.

Linking and merging data between more than two sources

In the previous section we described the linking between two sources of different data, one with survey respondents as units, the other with party manifestos as units. As illustrated in Figure 1, however, the PIREDEU data consist of five different components. Between each pair of these, the linking and merging of information proceeds along the logic described in the previous section, possibly with minor modifications necessitated by unique characteristics of each of these five components. When merging data from more than two sources we can distinguish two possibilities, one of which presents its own challenges. We use the notation [Primary (or recipient) & Secondary (or donor)] to denote the merging of information from the secondary into the primary dataset.

The easiest way to merge data between more than two sources consists of parallel merging, which is using the same source as primary dataset vis-a-vis all other sources as successive secondary datasets. In the case of the PIREDEU data, this could, for example, involve the MaS as primary data, into which information is merged, in successive rounds, from other sources such as the VS, the CS and the MS. In other words:

\[ \text{MaS} \, & \, \text{VS} \, + \, \text{MaS} \, & \, \text{CS} \, + \, \text{MaS} \, & \, \text{MS} . \]

The merging operation for each of these successive rounds is similar, and proceeds along the lines described in the previous section. In which order the successive rounds of merging are executed is immaterial for the final result. This example will result in a MaS with a great number of additional variables which contain information from the other data components.

A more complex way consists of sequential merging, where, for example, in a first round, VS data are merged into the MaS, while in the second, MaS information (including the data that have their origin in the VS) are merged into the:

\[ \text{MaS} \, & \, \text{VS} \, + \, \text{CS} \, & \, \text{MaS} \, & \, \text{VS}. \]

This implies that the identity of the primary dataset changes between successive rounds of merging. In these situations, the order of the successive rounds of merging is essential, because, as discussed earlier:

\[ \text{MaS} \, & \, \text{VS} \, \neq \, (\text{VS} \, & \, \text{MaS}). \]

The design of the user interface for multiple parallel merging is not intrinsically more complex than it is for single merging operations. However, in the case of the interface for multiple sequential merging, it is more difficult. The additional difficulty is not technical, but didactic in nature: the user interface is intended to allow users who are not used to RDB management to merge data from different components of a RDB by guiding them through a series of structured questions, the answers to which generate the syntax of the required RDB operations in the background. The challenge will thus be to implement possibilities for sequential merging, while keeping the interface simple and comprehensible.

Concluding remarks

We believe that our approach to producing 'integrated' datasets has a great advantage over other approaches. It does not invest in the production of a specific end-product, but rather in the tools to be used that will allow such a product to be achieved. The implication is that, once the tools are produced, different 'integrated' datasets can easily be produced from the same original empirical material. Or, stated differently, it does not produce a straightjacket to which researchers have to adapt themselves, but rather it allows the production of datasets that are tailored to the specific needs of individual researchers.

This approach to linking and merging data is in principle also applicable for other data than those collected in the context of the PIREDEU program. As a case in point, many studies that are conducted repeatedly – such as national election studies – strive to unleash the potential for longitudinal comparison by making available longitudinally integrated datafiles, often referred to as so-called continuity studies. But the construction of such datasets is never straightforward, as they invariably require many decisions being made to cope with unavoidable differences in operationalisations, coding schemes, and the like. Irrespective of what decisions are made, they can never be optimal for all the different research projects that would require such over-time comparable data. In these contexts too, it may be advantageous not to invest in the production of datafiles that will not be well suited for at least some researchers, but rather in a flexible interface that allows analysts to tailor the longitudinal data integration to the specific needs of their research.

References


Please address all correspondence to:

Methods and Data Institute
University of Nottingham
University Park, Law & Social Science Building
Nottingham NG7 2RD, UK

Notes

1. For full elaboration of the linking solution, see van der Eijk and Sapir (2010).

2. The website of the CSES lists a large number of comparative data collection projects in the general field of elections, parties and public opinion, with their respective URLs; see: http://www.umich.edu/~cses/about.htm.

3. Quite common, however, are publications in which separate and unconnected analyses from single datasets are ‘linked’ in a narrative fashion. Some of these are excellent and generate important insights. Yet, as will become obvious below, this nevertheless falls far short from linking the diverse data before the analysis stage and then analysing the merged data.

4. As a case in point, coding of parties is not fully comparable across successive editions of the European Social Survey (ESS), mainly by not anticipating that national party systems change over time. In 2008, code 11 for the variable asking about the party the respondent voted for in the last general elections in the Netherlands (PRVWNL) pertains to the PVV (Freedom Party), while in 2002 the same code pertains to ‘other party’ Such incomparabilities are particularly large in countries with unstable party systems.

5. We do not want to suggest that no useful efforts at harmonization are undertaken at all. Some of the most productive ones pertain to harmonization efforts aimed at making the coding of, e.g., educational attainments comparable across countries (the UNESCO initiated ISCED codes).

6. Detailed information about the PIREDEU program, including questionnaires and coding schemes, can be obtained from its website: http://www.piredeu.eu/.

As the focus of this paper is on data linking, we refrain here from presenting substantive information about the specific character of European Parliamentary elections, and we refer to the relevant literature, e.g., Reif and Schmitt (1980), van der Eijk and Franklin (1996), Schmitt and Thomassen (1999), van der Brug and van der Eijk (2007), and Thomassen (2009).

7. EES (2009); van Emgmond et al. (2010).

8. EES (2009a); Giebler, Haus & Weßels (2010).


10. EES (2009d); Schuck et al. (2010).

11. EES (2009b); Czesnik, Kotnarowski & Markowski (2010).

12. Such additional information can, of course, also be derived from external sources, such as the World Bank, OECD, EuroStat, and so on. For our example however we focus only on various PIREDEU data components.

13. This is similar to the futility of occasional proposals to ‘standardise’ the observation of essentially contested concepts (cf. Connolly 1999), or to standardise questionnaire items in survey research.

14. This implies that many of the attempts to provide ‘continuity’ files for, e.g., national election studies, are suboptimal at best. In the process of producing such files many operational decisions have to be made which are not of a technical and innocuous nature, but which have conceptual and theoretical implications. If analysts do not subscribe to these implications, the resulting continuity file will be less desirable, and they have to either repeat the same work on their own terms or, more frequently, abandon the project that required such linking.

15. For full elaboration of the linking solution, see van der Eijk and Sapir (2010).

16. Of course, many other summarising measures could also be relevant for interval level variables, such as x-tiles and x-tile ranges, skew and kurtosis, proportions above/below specified cut-off values, etc.
The International Association for Social Science Information Service and Technology (IASSIST) is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data.

Paid-up members enjoy voting rights benefit from reduced fees for attendance at regional and international conferences sponsored by IASSIST. Join today by filling in our online application:

http://www.iaassistdata.info/

Online Application

IASSIST Member ($50.00 (USD))
Subscription period: 1 year, on: July 1st
Automatic renewal: no

Please fill in the information our Online Form

The application is in USD, however, we do accept Canadian Dollars, Euro, and British Pounds as well.

The membership rates in all currencies as well as the Regional Treasurers who manage them are listed on the Treasurers page